Chinese word segmentation based on analogy and majority voting

Zongrong Zheng **Yi Wang** **Yves Lepage**
Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
{zzr0427@toki., yiwang@akane., yves.lepage@}waseda.jp

Abstract

This paper proposes a new method of Chinese word segmentation based on proportional analogy and majority voting. First, we introduce an analogy-based method for solving the word segmentation problem. Second, we show how to use majority voting to make the decision on where to segment. The preliminary results show that this approach compares well with other segmenters reported in previous studies. As an important and original feature, our method does not need any pretraining or lexical knowledge.

1 Introduction

Words are usually considered a basic unit in natural language processing (NLP) studies. As natural language texts are continuous sequences of characters, it is generally agreed that word segmentation is the initial step of NLP. The performance of the best Chinese segmenters for F-score has reached 95%, as reported in the second SIGHAN Chinese segmentation bakeoff (Emerson, 2005). These best existing methods rely on massive training data.

How to utilize as much information as possible from the training corpus to adapt a segmentation system towards a segmentation standard has been the main issue (Kit et al., 2005). Most of existing methods can be roughly classified as either dictionary-based or statistical-based methods.

Dictionary-based methods usually rely on large-scale lexicons and are built upon a few basic "mechanical" segmentation methods based on string matching. Without a large, comprehensive dictionary, the success of such methods degrade.

Statistical-based methods consider the segmentation problem as a classification problem on characters and usually involve complicated language models trained on large-scale corpora.

All of these methods require pre-training data and prior lexical knowledge. All current methods assume comprehensive lexical knowledge. How to model human cognition and acquisition it to segment words efficiently without using knowledge of wordhood is still a challenge in CWS (Huang et al., 2007).

After this introduction, we shall introduce the notion of proportional analogy in section 2 on which our proposal relies. In section 3, we shall describe the main idea of our new method for CWS using proportional analogy. Section 4 shall present the details of our implementation of our method. Section 5 shall detail some experiments done to evaluate our method with other state-of-the-art methods.

2 Proportional Analogy

Analogy has shown great potential in natural language processing, like machine translation (Lepage et al., 2005) and semantic relations (Turney et al., 2005). A proportional analogy is a relationship between four objects, noted $A : B :: C : D$ in its general form (Lepage et al., 2005). On numbers we have:

$$\frac{5}{15} = \frac{10}{30} \quad \text{also written as an analogy } 5 : 15 :: 10 : 30$$

By using words, sequences of words or sentences instead of numbers, we get proportional analogies

29th Pacific Asia Conference on Language, Information and Computation: Posters, pages 151 - 156
Shanghai, China, October 30 - November 1, 2015

between words, sequences of words or sentences. For instance, the following example is a true analogy between sequences of words:

$$I\ walked : to\ walk :: I\ laughed : to\ laugh$$

We use the algorithm proposed by Lepage (1998) for the resolution of analogical equations. This algorithm is based on the formalization of proportional analogies shown in formula (1) (Lepage, 2004).

$$A : B :: C : D \Leftrightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ dist(A,B) = dist(C,D) \\ dist(A,C) = dist(B,D) \end{cases} \quad (1)$$

Here, a is a character, whatever the writing system, and A, B, C and D are strings of characters. $|A|_a$ stands for the number of occurrences of character a in the string of characters A and $dist(A,B)$ stands for the edit distance between strings A and B which only considering insertions and deletions only as edit operations. The input of this algorithm is three strings of characters, words, sequences of words or sentences. Its output is a string of characters in analogy with the input. The following is an example applying this algorithm in Chinese:

我爱吃饭 : 我爱喝水 :: 你爱吃饭 : x

x = 你爱喝水

3 A New Method for CWS using proportional analogy

We propose a new Chinese word segmentation method based on proportional analogies. Crucially, we no longer need any pre-processing phase (training) or lexical knowledge (dictionary). The following gives the basic idea of the method. We are inspired by the example-based machine translation system proposed by Lepage et al. (2005).

Let us suppose that we have a corpus of sentences in their usual unsegmented form and their segmented form. We call it the training corpus. A line in such a training corpus may look like:

unsegmented form # segmented form
迈向充满希望的新世纪#迈向__充
满__希望__的__新__世纪

Let D be an input sentence to be segmented into segmented sentence $\tilde{D}$.

(i) We build all analogical equations $A_i : B_j :: x : D$ with the input sentence D and with all pairs of sub-strings (A_i, B_j) from the unsegmented part of the training corpus. According to formula (1), not all analogical equations have a solution. In order to get more analogical solutions and reduce time in solving analogical equations, we only consider sub-strings A_i and B_i which are more similar to D than a given threshold;

(ii) We gather all the solutions x of the previous analogical equations and only keep the solutions, named $C_{i,j}$, which belong to the training corpus. As it is easy to map from unsegmented part to segmented part for any sub-strings in training corpus, for each $C_{i,j}$, A_i and B_i, we easily retrieve their corresponding segmented forms$\widetilde{C_{i,j}}$, $\tilde{A}_i$ and $\tilde{B}_i$ in the segmented part of the training corpus;

(iii) We then form all possible analogical equations with all pairs $(A_i, B_j, C_{i,j})$:

$$\widetilde{A_i} : \widetilde{B_i} :: \widetilde{C_{i,j}} : y$$

We output the solutions $y = \widetilde{D_{i,j}}$ of all these analogical equations. They are hypotheses of segmentation for D. We record the number of times of each hypotheses. Recall that different analogical equations may generate identical solutions.

Figure 1 gives a simple example to illustrate the basic work flow of the method described above.

4 A CWS system using proportional analogy

In this section we describe the details of our implementation of the analogy-based word segmentation method. The key point in our method is to generate as precise proportional analogies as possible. These solutions of proportional analogy are the segmented results of input sentences. As not all of these solutions are exactly correct, we will consider them

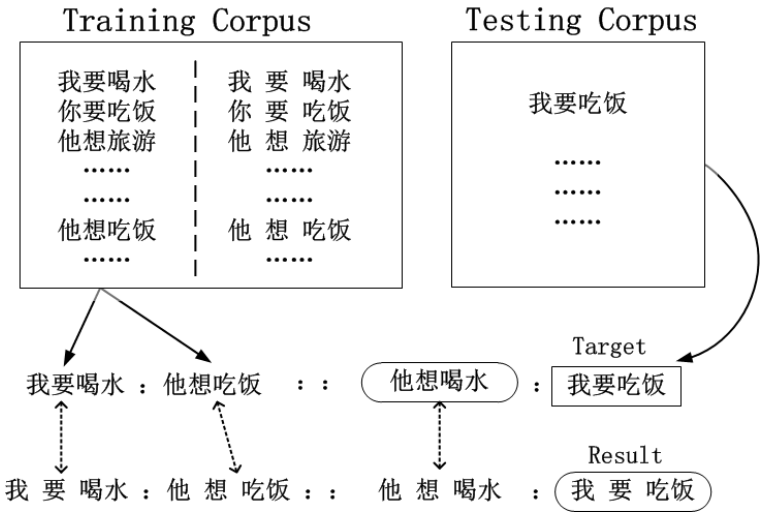

Figure 1: Illustration of the Chinese word segmentation method based on proportional analogy

as hypotheses of segmentation. According to formula (1), the longer the sentences are, the more difficult the constrained equations are satisfied. It means that long sentences are easy to miss analogical solutions and further to miss hypotheses of segmentation. Splitting sentences is necessary. We split sentences into n-grams, i.e., sub-strings of length n. Our system is thus divided into two parts: generating hypotheses of segmentation for n-grams and recombining strategy for segmentation hypotheses to generate a complete segmented result for the entire input sentences.

4.1 Generating segmented references of n-grams

We adopt the method proposed in section 3 to generate the segmented result of n-grams in practice in our system. The work flow of generating a segmentation hypotheses for n-grams is shown in figure 2.

According to formula (1), A and B should share characters with D to get a solution from equation $A_i : B_j :: x : D$. It means that A and B should be similar strings to D to a certain extent. We use TRE agrep[1], an approximate regex matching library, to retrieve sub-strings which are similar to the input D from training corpus. We use edit distance, with only insertions and deletions as edit operations, to quantify how similar two strings are to one another. Any two of these similar substrings and input D form an analogical equation. In general, not all solutions of the equations occur in the training corpus. Consequently, only the solutions which occur in the segmented part of the training corpus are considered as segmentation hypotheses. Notice that different analogical equations may generate identical solutions. The same segmentation hypotheses can be generated several times by different analogical equations. We record this number of occurrences. It is natural to think that the larger the number of occurrence is, the more likely the segmentation hypothesis is.

[1] http://laurikari.net/tre/

4.2 Recombination Strategy

We use majority voting rules to recombine the segmentation hypotheses of n-grams. A segmentation hypothesis can be represented as a sequence of characters and delimiters. The general form is:

$$c_1 D_1 c_2 D_2 ... c_{n-1} D_{n-1} c_n,$$

$$occurrence\ number = m.$$

In this form, D_i is either a space or not a space. We let all segmentation hypotheses vote for D_i.

When D_i is a space, it means that this segmentation hypothesis votes m times for segmentation. When D_i is not a space, it votes m times against segmentation. Figure 3 is an example to illustrate the use of majority voting in our system. We sum

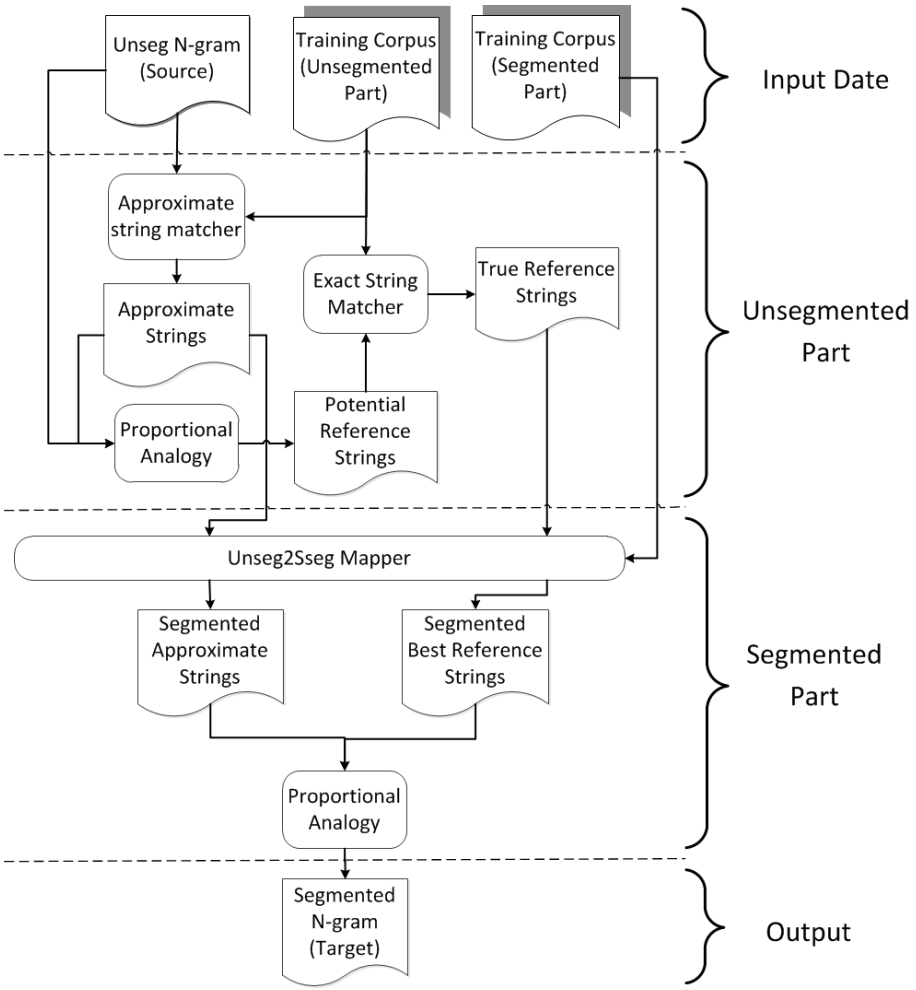

Figure 2: Work flow of generating segmented reference of n-grams in our system

	c_1	D_1	c_2	D_2	c_3	D_3	c_4	D_4	c_5	D_5	c_6	D_6	c_7	D_7	c_8	D_8	c_9
# of occurrence	人		类		社		会		前		进		的		航		船
9	人		类	␣	社		会	␣	前								
3			类	␣	社		会	␣	前		进						
1					社		会		前	␣	进	␣	的				
11					社		会		前		进	␣	的				
37					社		会	␣	前		进	␣	的				
4									前		进	␣	的	␣	航		船
1									前		进		的		航		船
for seg (␣)		0		**12**		0		**49**		1		**53**		**4**		0	
against seg		**9**		0		**61**		12		**56**		1		1		**5**	
segmentation result	人		类	␣	社		会	␣	前		进	␣	的	␣	航		船

Figure 3: An example of recombination of segmentation hypotheses of n-grams using majority voting

	PKU
Word tokens	104372
Word types	13148
OOV words tokens	6006
OOV words types	2863
Character tokens	172733
Character types	2934
OOV character tokens	372
OOV character types	92

Table 1: Corpus details of PKU test set.

up the votes in favor and against segmentation and output the final results according to the vote results.

5 Experiments

5.1 Data and Evaluation

To evaluate the effectiveness of our proposed method, we conduct experiments on a widely used Chinese word-segmented corpora, namely PKU, from the second SIGHAN international Chinese word segmentation bakeoff (Emerson, 2005). The training set and the test set are publicly available from the official website[2]. Table 1 shows some statistics on the data sets. All evaluation results in this paper are tested by the official scoring script, also downloaded from the official website.

The segmentation accuracy is evaluated by test recall (R), test precision (P) and balanced F-score, as defines in Equation (2), (3) and (4).

$$\text{R} = \frac{\text{number of correctly segmented words}}{\text{total number of words in gold standard segmentation}} \quad (2)$$

$$\text{P} = \frac{\text{number of correctly segmented words}}{\text{total number of words in segmentation result}} \quad (3)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (4)$$

Our experiments follow the closed track. It means that no extra resource other than training corpora is used.

[2]http://www.sighan.org/bakeoff2005/

Models	PKU Corpus				
	P	R	F	R_{oov}	R_{iv}
baseline	84.3	90.7	87.4	6.9	95.8
Best05 closed-set	95.4	94.6	95.0	78.7	95.6
This work (closed-set)	90.9	89.9	90.4	60.7	91.6

Table 3: Performance of our system on the SIGHAN 2005 data set. Best05 refers to the best closed-set results in SIGHAN 2005 bakeoff.

5.2 Effects of Length of n-grams and Edit Distance

As discussed in section 4, long sentences are easier to miss hypotheses of segmentation. So the length of n-grams will influence the segmentation results. Moreover, the larger edit distance is used, the more similar sub-strings would be retrieved. To measure it, we conduct experiments using different length of n-grams and different edit distance.

According to our majority voting method, we would consider a position is not segmented if no segmentation hypothesis votes for it. The results in Table 2 shows that this data sparse problem is more serious when we used larger length of n-grams.

5.3 Results

We set length of n-grams to 3 and edit distance to 2 for approximate string match to perform our experiments. Table 3 shows our empirical results on the data set. Our system achieve a significantly better results than the baseline. R_{iv} score shows that our method performs well on in vocabulary (IV) word recognition. Simultaneously, the R_{oov} score shows that our method has certain ability to deal with out-of-vocabulary (OOV) word and guess their form. Compared with best result (Tseng et al., 2005) in SIGHAN 2005, our result still has a lot of room for improvement. But as a original method which do not need any pre-training or lexical knowledge, our method has a great potential in CWS.

6 Conclusion

In this paper, we presented an approach to Chinese word segmentation based on proportional analogy and majority voting to make decision on where to segment. Our approach achieves a desirable accuracy, when evaluated on the corpus of the close track of SIGHAN 2005 and shows an excellent perfor-

# of n-grams	Edit Distance	Word Count	P	R	F
6	3	79828	85.5	65.4	74.1
5	3	95079	90.0	82.0	85.8
4	2	99103	90.8	86.2	88.4
3	2	103186	90.9	89.9	90.4

Table 2: Performance of our method with different length of n-grams and edit distance.

mance in word identification. As an important and original feature, our method does not need any pre-training or lexical knowledge.

References

Thomas Emerson. The second international chinese word segmentation bakeoff. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 133, 2005.

Chu-Ren Huang, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot. Rethinking chinese word segmentation: tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 69–72. Association for Computational Linguistics, 2007.

Chunyu Kit and Xiaoyue Liu. An example-based chinese word segmentation system for cwsb-2. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 146–149, 2005.

Yves Lepage. Solving analogies on words: an algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 728–734. Association for Computational Linguistics, 1998.

Yves Lepage. Analogy and formal languages. *Electronic notes in theoretical computer science*, 53:180–191, 2004.

Yves Lepage and Etienne Denoual. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282, 2005.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171, 2005.

Peter D Turney and Michael L Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278, 2005.